

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-212178
(43)Date of publication of application : 20.08.1996

(51)Int.Cl.

G06F 15/163
G06F 13/28

(21)Application number : 07-020137
(22)Date of filing : 08.02.1995

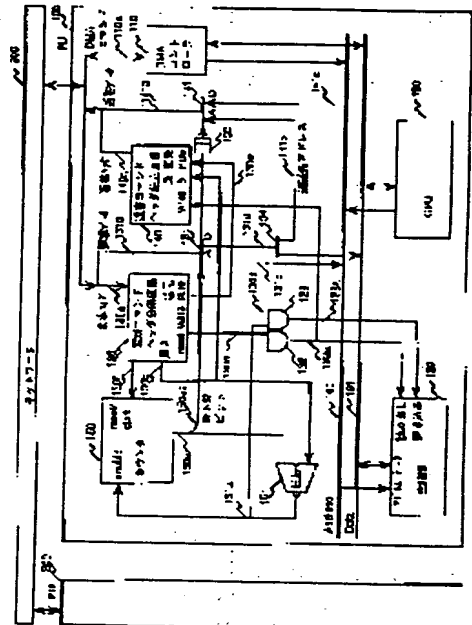
(71)Applicant : HITACHI LTD
(72)Inventor : TARUI TOSHIAKI
AKASHI HIDEYA
SUKEGAWA NAONOBU
FUJII KEIMEI

(54) PARALLEL COMPUTER

(57)Abstract:

PURPOSE: To reduce the overhead of accesses by issuing a single network command for the requests of accesses to be given to plural discontinuous data on the addresses of other processing units (PU) and then having the automatic accesses to these data by means of hardware.

CONSTITUTION: In a write command mode, the addresses included in the even-numbered words are outputted to an address bus 160 and the data included in the odd-numbered words are outputted to a data bus 161. Then these addresses and data are written in a main storage 120. Thereby, the writing operations can be requested by a single network command to the discontinuous addresses of the main storage 120. In a read command mode, the addresses included in the even-numbered words are outputted to the bus 160 and an access is given to the storage 120. Then these addresses are paired with the return destination addresses included in the odd-numbered words of a request command, and a remote write command is acquired by a selector 141 and a header assembly circuit 140. This write command is sent back to a requester PU which processes the received command as a write command.



LEGAL STATUS

[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-212178

(43) 公開日 平成8年(1996)8月20日

(51) Int.Cl.⁶

G 0 6 F 15/163

13/28

識別記号

庁内整理番号

F I

技術表示箇所

3 1 0 C 9172-5E

G 0 6 F 15/ 16

3 2 0 Z

審査請求 未請求 請求項の数 8 O L (全 7 頁)

(21) 出願番号

特願平7-20137

(22) 出願日

平成7年(1995)2月8日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 垂井 俊明

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 明石 英也

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 助川 直伸

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 弁理士 小川 勝男

最終頁に続く

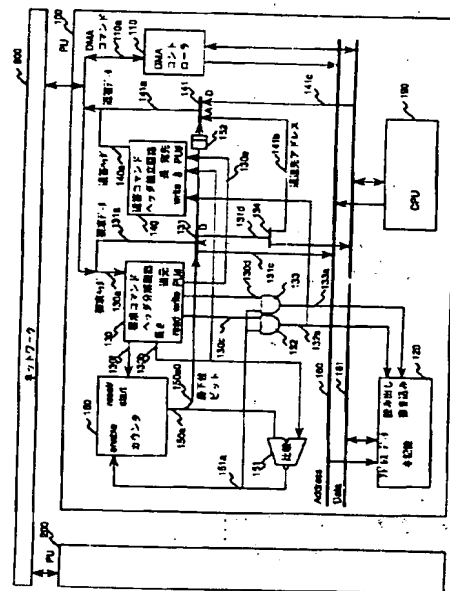
(54) 【発明の名称】 並列計算機

(57) 【要約】

【目的】 分散記憶を持った並列計算機において、他のプロセッシングユニット (P U) にある、アドレスの連続しない複数のデータをアクセスする際の待ち時間、オーバヘッドを低減する。

【構成】 他 P U への書き込み処理では、ネットワーク上のコマンドで、書き込むデータ 1 ワード毎に書き込みアドレスを指定する。コマンドを受け取った P U では、ネットワークコマンド中の、アクセスアドレス、データの組をアドレスバス、データバスに振り分け、主記憶に書き込む。他 P U データの読み出し処理も、ネットワーク上のコマンドで、読み出すデータ 1 ワード毎に、読み出しアドレスと、読み出したデータを格納するための返送先アドレスを指定する。コマンドを受け取った P U では、各々のアドレスのデータを読み出し、返送先アドレスに返送する。

図 1



【特許請求の範囲】

【請求項1】複数のプロセッシングユニットを持ち、各プロセッシングユニットが独立した主記憶を持ち、前記各プロセッシングユニットがネットワークにより接続されている並列計算機において、他のプロセッシングユニットの主記憶上にある、アドレスが連続しない複数のデータに対するアクセスを、一つのネットワークコマンドで指定することを特徴とする並列計算機。

【請求項2】請求項1において、任意のアドレスを持つ複数のデータへの書き込みを、同一のネットワークコマンドで要求する並列計算機。

【請求項3】請求項2において、前記ネットワーク上のデータ書き込みコマンドで、書き込みアドレス、書き込みデータの組を任意の個数持つことが可能である並列計算機。

【請求項4】請求項3において、他プロセッシングユニットから到来した書き込みコマンド中の、書き込みアドレス、書き込みデータを、主記憶のアドレス線、データ線に振り分けるためのスイッチを持ち、主記憶への書き込みを行う並列計算機。

【請求項5】請求項1において、任意のアドレスを持つ複数のデータへの読み出しを、同一のネットワークコマンドで要求する並列計算機。

【請求項6】請求項5において、他のプロセッシングユニットから読み出した複数のデータを、自プロセッシングユニットの主記憶上の任意の位置に置くことが出来る並列計算機。

【請求項7】請求項6において、ネットワーク上のデータ読み出しコマンドで、他プロセッシングの主記憶上の読み出しアドレス、読み出したデータを格納する自プロセッシングユニット上のアドレスの組を任意の個数持つことが可能である並列計算機。

【請求項8】請求項7において、他プロセッシングユニットから到来した読み出しコマンド中の、読み出しアドレス、読み出したデータを格納するアドレスを、主記憶のアドレス線、読み出したデータの返送先アドレスに振り分けるためのスイッチを持ち、返送先アドレス、読み出されたデータの組複数個を、一つの書き込みコマンドにまとめ、ネットワークへ出力するためのセクタを持つ並列計算機。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は複数のプロセッシングユニットからなる並列計算機におけるデータ転送方式に関する。

【0002】

【従来の技術】計算機性能の飛躍的向上に関して、多数台のプロセッシングユニット（以下、PU）を並列動作させる、並列計算機が有望視されている。並列計算機では、多数台のPUの間で効率良くデータを通信すること

が重要で、特に大規模な数値演算では、計算に必要な大量のデータを、PU間で一括して高速に転送するためのアーキテクチャが必要である。

【0003】従来の並列計算機におけるデータ転送機構は、特開平6-19856号公報に示されているように、連続したアドレスのデータを一括して転送する機構が採用されていた。各PUは他PUの主記憶との間でDMAを行うための機構を持ち、転送したいデータの領域を指定すると、DMA機構のハードウェアが指定された領域を自動的に転送する。

【0004】

【発明が解決しようとする課題】上記従来技術では、転送しようとするデータが全て連続したアドレス（もしくはストライドアクセスなどの定型的なパターン）に存在する場合は有効であるが、転送するデータのアドレスが、連続でないランダムなアドレスの場合には、効率が悪いという問題がある。

【0005】例えば、リモートの非連続な領域に書き込みを行う場合を考える。その場合、従来の連続アドレスへの書き込みのみが可能なDMAデータ転送機構では、相手のメモリ上に複数のデータを一括して書き込むことができない。そのため、

(1) 1ワード毎にリモート書き込みコマンドを出す。

(2) 書き込むデータと書き込むアドレスを入れた2本の配列を、一旦、相手先PUの別々の領域に2回に分けて転送した後、相手先のPUに、本来書き込むべき領域への実際の書き込み処理を依頼する。等の方式が取られていた。

【0006】(1)の方式は、1ワード毎の書き込みコマンドを多数送出しなければならないため、実行時間が増大するばかりか、ネットワーク上に大量のバケットを出す必要が有るため、ネットワークの負荷が増大し、問題である。

【0007】(2)の方式は、ネットワークの負荷は軽減されるが、一旦アドレス、データを着地させる領域が新たに必要になり、メモリの使用効率が落ちる。さらに、相手先のPUに余分な仕事が発生するため、プログラムの実行時間が増大する問題がある。

【0008】リモートのPUの非連続なアドレスにあるデータを読み出そうとした場合も、

(1) 1ワード毎に読み出す

(2) 相手先のCPUに依頼してデータを連続領域に集めてもらった後、一括して転送する
等の処理が必要となり、書き込みの場合と同様に処理効率の大幅な低下を招く。

【0009】特に近年、非定型なデータを扱うため、データ構造としてリストベクトルを利用したプログラムが数多く見られる。リストベクトルのアクセスの場合、アクセス先のアドレスはポインタの配列で示されているため、一般には非連続なアドレスへのデータアクセスと

なる。従ってリストベクトルのプログラムを高速に実行するには、非連続なアドレスにあるデータを一括して高速に転送する機構が必要となる。

【0010】

【課題を解決するための手段】上記目的を達成するために、複数のデータを転送するためのネットワークコマンドの中で、アクセスを行うリモートアドレスを1データ毎に指定することが可能なコマンド構造とする。

【0011】他PUの非連続なアドレスへの書き込みの場合、書き込み先のPUへ送られるネットワークコマンドの中に、書き込みアドレスと書き込むべきデータの組を、任意の個数持たせる。上記コマンドを受け取ったPUは、コマンド中の各組のアドレス部分に入っている値をアドレスバスに、データ部分に入っている値をデータバスに振り分け、アドレスで示される領域にデータを書き込む処理をコマンドの長さだけ繰り返す。これにより、他PUのアドレスが非連続な複数のワードへの書き込みを一つのコマンドで指示することができる。

【0012】非連続なアドレスの読み出しの場合には、ネットワーク上の要求コマンドに読み出しアドレスと、読み出したデータを書き込むべき要求側のPUの主記憶のアドレス（以下では返送先アドレスと呼ぶ）の組を、任意の個数持たせる。上記コマンドを受け取ったPUは、コマンド中の各組の中のアドレス部分に入っている値をアドレスバスに出力して主記憶中の値を読み出した後、返送先アドレスに書き込むための処理を行う。ここで、読み出した値の返送先アドレスへの書き込みは、それ自体、複数の非連続なアドレスへの書き込みとなるので、前に記した非連続なアドレスへの書き込みコマンドを用いて、読み出した値を返送先アドレスへ書き込むように指示する。これにより、他PUの、アドレスが非連続な複数のワードのデータを読み出し、自PUの領域に書き込むことが出来る。

【0013】

【作用】本発明によれば、ネットワーク上を流れる主記憶アクセスコマンドで、アクセス先のアドレスをデータ毎に指定し、さらに、返答側PUのハードウェアで前記コマンドを分解し、主記憶をアクセスするハードウェアを設ける。これにより、他PUの主記憶上のアドレスが連続しない複数のデータを、1回のネットワークコマンドで高速にアクセスすることができる。

【0014】図1に本発明の並列計算機のブロック図を示す。図中130が他PUからの要求バケットのヘッダを分解するための回路、131が要求バケットのデータ部のアドレスとデータ等を振り分けるための回路である。データ部のワードの数をカウンタ150で数える。書き込みコマンドの場合、偶数ワードに入っているアドレスはアドレスバス160へ、奇数ワードに入っているデータはデータバス161へ出力し、主記憶への書き込みを行う。これにより、主記憶の非連続なアドレスへの書

き込みを一つのネットワークコマンドで依頼することができる。

【0015】読み出しコマンドの場合は、偶数ワードに入っているアドレスをアドレスバス160に出力し、主記憶をアクセスした後、要求コマンドの奇数ワードに入っている返送先アドレスと組にして、セクタ141およびヘッダ組立回路140でリモート書き込みコマンドを組立て、要求元PUに返送する。要求元PUでは前述の書き込みコマンドとして処理を行うことにより、他PUの非連続なアドレスにあるデータを、自PUの非連続なアドレスに転送することができる。

【0016】図中のコンパレータ151はコマンド中のアクセスワード数とカウンタ150の値を比べ、コマンド処理の終了を検出するための回路である。これにより、バケット中でアクセスするワード数を任意に指定でき、柔軟なリモートアクセスを行うことが出来る。

【0017】

【実施例】図1ないし図4に本発明の一実施例を示す。図1は本発明の並列計算機のブロック図である。図2ないし図4はPU間ネットワークのコマンドバケットのフォーマットである。図2は非連続なアドレスへ複数のデータの書き込みを指定するためのコマンド（以下ではマルチワードライトと呼ぶ）、図3は非連続なアドレスの複数のデータの読み出しを指定するためのコマンド（以下ではマルチワードリードと呼ぶ）である。それに対して、図4は従来のDMA書き込みのコマンド（以下ではDMAライトと呼ぶ）である。

【0018】図1において、100、200はPU、900はPU間ネットワークである。以下ではPU100の内部のみ詳細に記す。他のPUも全く同一の構成を持つ。PUの内部では、190がCPU、120が主記憶、160がアドレスバス、161がデータバス、110がPU間で従来型のDMA転送を行うためのDMAコントローラである。さらに、130は他PUからのマルチワードリード、マルチワードライトコマンドのヘッダ部を解釈するための要求コマンドヘッダ分解回路、131は要求コマンドのデータ部のアクセスアドレスと、書き込みデータ（書き込みの場合）又は返送先アドレス（読み出しの場合）を振り分けるためのスイッチである。150はデータ部のワード数を数えるためのカウンタ、150a0はカウンタの最下位ビット、151はコマンドバケットの終了を判定するための比較器、132、133は主記憶アクセスコマンドを出力するための回路である。134はマルチワードリード、マルチワードライトコマンドを切り替えるためのスイッチである。140はマルチワードリードの返答を行うためのマルチワードライトコマンドのヘッダを組み立てる回路、141はマルチワードリードの返答を行うためのマルチワードライトコマンドのデータ部の、返送先アドレスとデータの組を組み立てるためのセクタである。

【0019】本発明では、各PUがマルチワードコマンドを実行するために、バケット中のアドレス情報をスイッチ131により切り分け、主記憶をアクセスする機構を持つことに特徴がある。

【0020】先ず、システム全体の構成について述べる。システムは、プログラムを実行するPU(100, 200)が、ネットワークにより接続された構成を取る。各PUはCPU190及び主記憶120を持ち、主記憶分散型のマルチプロセッサシステムを構成している。PU間の通信はネットワークを経由したバケット通信で行われる。

【0021】通常(従来型の)PU間の通信は110のDMA通信機構によって主記憶上のあるまとまった領域を一括して転送することにより行われる。

【0022】図4にネットワーク上のDMAライトコマンドのフォーマットを示す。ネットワークコマンドはヘッダとしてコマンド名1001、宛先PU番号1002、コマンド長(データ部のワード数)1003、送信元PU番号1004が置かれる。ヘッダ部の後にデータ部が置かれる。データ部ではDMAの送り先アドレス1300aに引続き、DMAで送られるデータ1300d~1306dが置かれる。ここで、この実施例でのデータの1ワードは4Bである。DMAの送信側では、図4に示されるバケットをCPUが主記憶上に作成し、DMAコントローラは主記憶上のバケットをネットワークに転送する。DMAライトコマンドを受けたPUのDMAコントローラは、DATA0~DATA $n-1$ のデータを開始アドレスで示される領域から順番に書き込む。DMAコントローラの詳細については既知の技術であるのでここでは説明を略す。

【0023】次にマルチワードライトコマンドの動作について述べる。図2にマルチワードライトコマンドのフォーマットを示す。ヘッダ部はDMA転送と同じであるが、データ部の形式が異なる。DMA転送ではデータ部で指定される転送先アドレスは一つであるのに対し、マルチワードライトではデータ1ワード毎にアドレスが指定される。図の例では、Addr0で示されるアドレスにData0を、Addr1にData1を、という様に、各データを別々のアドレスに書き込むことができる。

【0024】マルチワードライトの要求側のPUでは、図2のバケットをCPUが予め主記憶に作成し、DMAコントローラ110を利用して主記憶上のバケットをネットワークに転送する(この部分はDMAライトと全く同じである)。

【0025】マルチワードライトを受信したPUはバケットのヘッダを要求コマンドヘッダ分解回路130に、データ部をスイッチ131に送る。要求コマンドヘッダ分解回路では、バケットのヘッダ部を分解し、コマンド種に応じてマルチワードライトの場合は信号130dを出力すると同時に、データ部の長さ130b、送り元PU番号130eを出力する。さらにワード数カウンタ

150にリセット/スタート信号130fを送る。カウンタ150の出力とバケット中の長さフィールド130bは比較機151で比較され、両者の値が異なる(カウンタの値がバケットの長さより小さい)間、信号151aが出力される。151aにより、カウンタ150がイネーブルされると同時に、ゲート133によって、書き込みコマンド133aが主記憶120に伝えられる。

【0026】スイッチ131は、カウンタの最下位ビット150a0の値、つまり、バケットのデータ部の偶数ワードか奇数ワードかに応じ、バケットのデータ部の値131aを、アドレス131c(偶数ワードの場合)とデータ131d(奇数ワードの場合)に振り分ける(スイッチ134はマルチワードライトの場合データバス161に接続されてる)。これにより、バケット中のアドレスとデータの組をアドレスバス160とデータバス161に出力し、主記憶120に書き込むことが出来る。

【0027】カウンタ150の値がバケットのデータ長130bと等しくなる(バケットが終了する)と、151a信号が出力されなくなる。それにより、主記憶への書き込み信号133aが止められ、カウンタ150の動作が止められ、処理が終了する。

【0028】以上の処理により、マルチワードライトコマンドの中の各アドレス、データの組を主記憶に書き込むことが出来る。

【0029】次に、マルチワードリードの動作について述べる。図3にマルチワードリードコマンドのフォーマットを示す。ヘッダ部はDMA転送等と同じである。マルチワードリードは、相手先PUの任意のアドレスの値を読み、自PUの任意のアドレスに書き込むためのコマンドである。データ部には読み出す相手先PUのアドレスと、読み出したデータを書き込む自PU上のアドレス(返送先アドレス)の組を複数持つ。

【0030】図3の例では、相手先PUのAddr0で示されるアドレスのデータを読み出し、自PUのDest0で示されるアドレスに書き込み、Addr1のデータをDest1に、という様に、相手先PUの別々のアドレスのデータを読み出し、自PUの別々のアドレスに書き込むことが出来る。

【0031】マルチワードリードの要求側のPUでは、図3のバケットをCPUが予め主記憶に作成し、DMAコントローラ110を利用して主記憶上のバケットをネットワークに転送する。

【0032】マルチワードリードを受信したPUはバケットのヘッダを要求コマンドヘッダ分解回路130に、データ部をスイッチ131に送る。要求コマンドヘッダ分解回路では、バケットのヘッダ部を分解し、コマンド種に応じて、マルチワードリードの場合は信号130cを出力すると同時に、データ部の長さ130b、送り元PU番号130eを出力する。さらにワード数カウンタ

150にリセット/スタート信号130fを送る。カウンタ150の出力とバケット中の長さフィールド130bは比較機151で比較され、両者の値が異なる(カウンタの値がバケットの長さより小さい)間、信号151aが出力される。151aにより、カウンタ150がイネーブルされると同時に、ゲート132によって、読み出しコマンド132aが主記憶120に伝えられる。

【0033】スイッチ131は、カウンタの最下位ビット150a0の値、つまり、バケットのデータ部の偶数ワードか奇数ワードかに応じ、バケットのデータ部の値131aを、アクセスアドレス131c(偶数ワードの場合)と返送先アドレス141b(奇数ワードの場合)に振り分ける(スイッチ134はマルチワードリードの場合セクタ141に接続されてる)。その後、アドレスバス160上のアドレスを用いて、主記憶の値が読み出され、読み出されたデータは、データバス161を通り、セクタ141に入力される。

【0034】その後、返答コマンドヘッダ組立回路140、セクタ141を用いて、読み出された値を送り元のPUに返送するためのマルチワードライトコマンドが出力される。このコマンドはAddr0~Addrn-1に格納されていた値を、送り元PUのDest0~Destn-1に書き込む。

【0035】まず、返答コマンド組立回路140は要求コマンドヘッダ分解回路130から伝えられた送り元PU(つまり返答コマンド宛先PU)番号130e、コマンド長130bより、返答用のマルチワードライトコマンドのヘッダを送出する。バケットのデータ部141aには、スイッチ141を用いて、カウンタの最下位ビット150a0の値(ただし、返答回路では、主記憶アクセスを待つために、ディレイラッチ152を用いて1サイクル遅らせてある)、つまり、バケットのデータ部が偶数ワードか奇数ワードかに応じ、返送先アドレス141b(偶数ワードの場合)もしくは読み出したデータ141c(奇数ワードの場合)を出力する。これにより、送り元PUから送られてきた返送先アドレスと、主記憶を読み出したデータの組をマルチワードライトコマンドのデータ部として送り元PUに返送することができる。

【0036】カウンタ150の値がバケットのデータ長130bと等しくなる(バケットが終了する)と、151a信号の出力が止められる。それにより、主記憶への読み出し信号132aが止められ、カウンタ150の動作が止められ、返送コマンドの送出も終了する。以上の

処理により、マルチワードリードコマンドの中の各アドレスの値を読み出し、送り元PUの返送先アドレスに書き込むことができる。

【0037】以上の方式により、マルチワードライト、マルチワードリードコマンドを用いて、他PUの主記憶上のアドレスが非連続な複数のワードに対する、書き込み、読み出し処理を一つのネットワークコマンドで一括して行うことが可能である。

【0038】

10 【発明の効果】本発明によれば、分散メモリ型の並列計算機において、他PUにある、アドレスの連続しない複数のデータに対するアクセスを、一つのネットワークコマンドで依頼し、ハードウェアで自動的に行うことにより、従来の連続アドレスに対するアクセスのみが可能なDMA機構を使用した場合と比較して、アクセスのオーバーヘッドを大幅に削減することが可能になる。

【図面の簡単な説明】

【図1】本発明の一実施例のリモートアクセス機構を持った並列計算機のブロック図。

20 【図2】ネットワーク上のマルチワードライトコマンドのフォーマットを示す図。

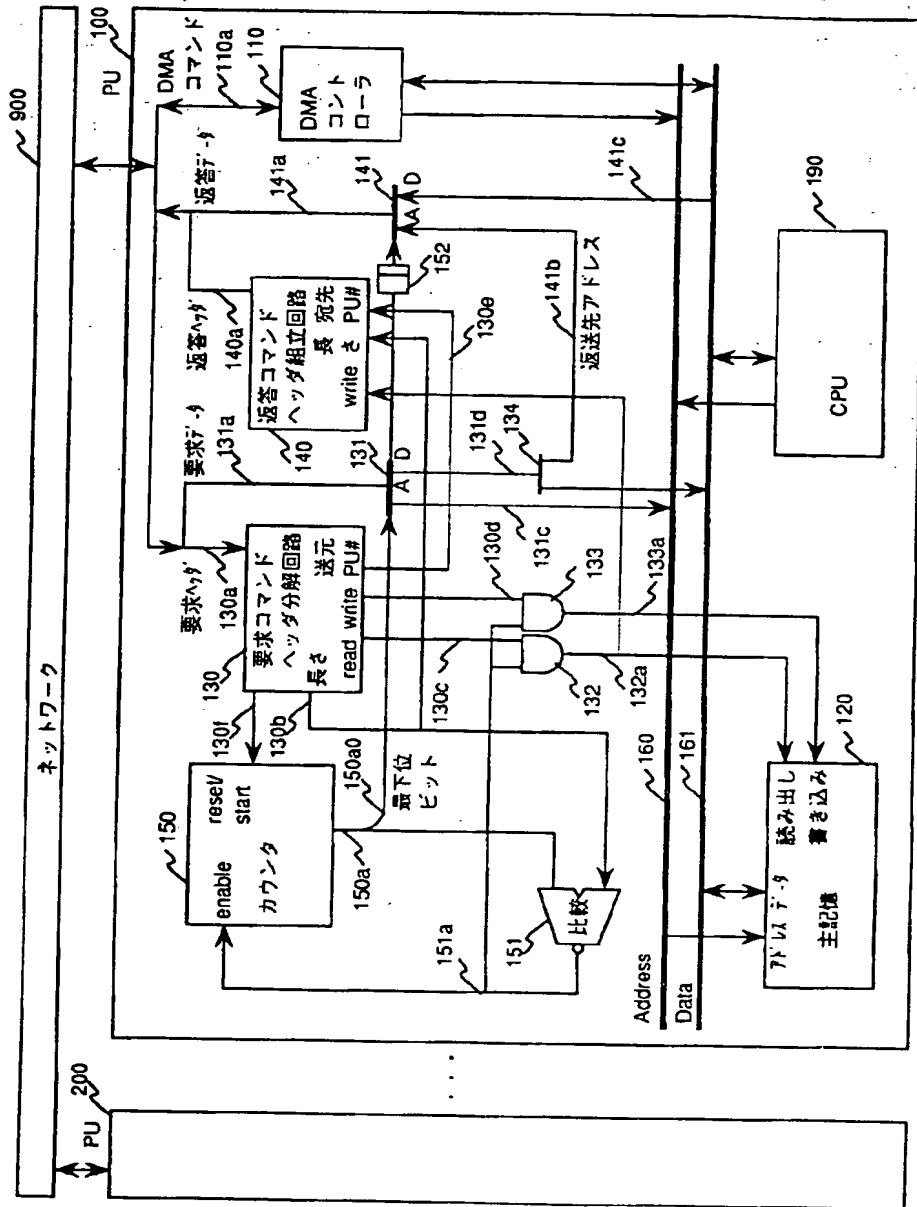
【図3】ネットワーク上のマルチワードリードコマンドのフォーマットを示す図。

【図4】従来のDMA書き込みコマンドのフォーマットを示す図。

【符号の説明】

100、200…プロセッシングユニット、110…DMAコントローラ、110a…DMAコマンド、120…主記憶、130…分解回路、130a…要求コマンドヘッダ、130b…データ長、130c…マルチワードリード信号、130d…マルチワードライト信号、130e…送り元PU番号、130f…カウンタコントロール信号、131…スイッチ、131a…要求コマンドデータ、131c…主記憶アドレス、131d…返送先アドレス、132…信号出力ゲート、133…信号出力ゲート、134…切替スイッチ、140…組立回路、140a…返答コマンドヘッダ、141…セクタ、141a…データ、141b…返送先アドレス、141c…読み出しデータ、150…ワード数カウンタ、150a…カウンタ出力、150a0…カウンタ出力最下位ビット、151…コンパレータ、151a…コマンドイネーブル信号、152…ラッチ、160…アドレスバス、161…データバス、190…CPU、900…ネットワーク。

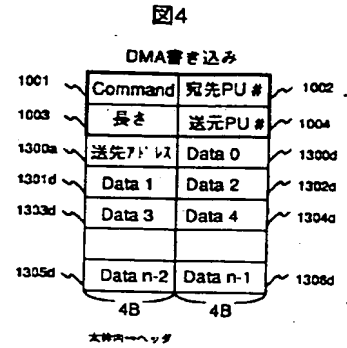
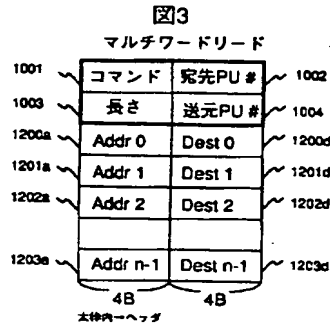
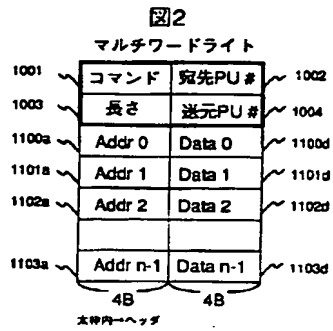
图 1



【図2】

【図3】

【図4】



フロントページの続き

(72)発明者 藤井 啓明

東京都国分寺市東恋ヶ窪1丁目280番地
株式会社日立製作所中央研究所内